2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Localizing and Quantifying Damage in Social Media Images

Xukun Li
Department of Computer Science
Kansas State University
Manhattan, Kansas
Email: xukun@ksu.edu

Huaiyu Zhang
Department of Statistics
Kansas State University
Manhattan, Kansas
Email: huaiyu@ksu.edu

Doina Caragea
Department of Computer Science
Kansas State University
Manhattan, Kansas
Email: dcaragea@ksu.edu

Muhammad Imran
Qatar Computing Research Institute
Hamad Bin Khalifa University
Doha, Qatar
Email: mimran@hbku.edu.qa

***Abstract*—Traditional post-disaster assessment of damage heavily relies on expensive GIS data, especially remote sensing image data. In recent years, social media has become a rich source of disaster information that may be useful in assessing damage at a lower cost. Such information includes text (e.g., tweets) or images posted by eyewitnesses of a disaster. Most of the existing research explores the use of text in identifying situational awareness information useful for disaster response teams. The use of social media images to assess disaster damage is limited. In this paper, we propose a novel approach, based on convolutional neural networks and class activation maps, to locate damage in a disaster image and to quantify the degree of the damage. Our proposed approach enables the use of social network images for post-disaster damage assessment, and provides an inexpensive and feasible alternative to the more expensive GIS approach.**

***Index Terms*—image analysis, convolutional neural networks (CNN), class activation mapping (CAM), damage localization**

I. Introduction

Fast detection of damaged areas after an emergency event can inform responders and aid agencies, support logistics involved in relief operations, accelerate real-time response, and guide the allocation of resources. Most of the existing studies on detecting and assessing disaster damage rely heavily on macro-level images, such as remote sensing imageries [1], [2], [3], or imageries transmitted by unmanned aerial vehicles [4]. Collection and analysis of macro-level images require costly resources, including expensive equipment, complex data processing tools, and also good weather conditions. To benefit the response teams, the macro-level images have to be collected and analyzed very fast, which is not always possible with traditional collection and analysis methods.

With the growth of social media platforms in recent years, real-time disaster-related information is readily available, in the form of network activity (e.g., number of active users, number of messages posted), text (e.g., tweets), and images posted by eyewitnesses of disasters on platforms such as Twitter, Facebook, Instagram or Flicker. Many studies have shown the utility of social media information for disaster management and response teams. For example, the analysis of text data (e.g., tweets from Twitter) has received significant attention in recent works [5], [6], [7], [8], [9]. However, social media images, while very informative [10], have not been extensively used to aid disaster response, primarily due to the complexity of information extraction from (noisy) images, as compared to information extraction from text.

By contrast with macro-level images, social media images have higher "resolution", in the sense that they can provide detailed on-site information from the perspective of the eye-witnesses of the disaster [10]. Thus, social media images can serve as an ancillary yet rich source of visual information in disaster damage assessment. Pioneering works with focus on the utility of social media images in disaster response include [11], [12], where the goal is to use convolutional neural networks (CNN) to assess the severity of the damage (specifically, to classify social media images based on the degree of the damage as: *severe*, *mild*, and *none*).

Due to ground-breaking developments in computer vision, many image analysis tasks have become possible. Disaster management and response teams can benefit from novel image analyses that can produce quantitative assessments of damage, and inform the relief operations with respect to priority areas. In this context, it would be useful to locate damage areas in social media images (when images contain damage), and subsequently use the identified damage areas to assess the damage severity on a continuous scale. Possible approaches for localizing damage in social media images include object detection [13], [14] and image segmentation [15]. Object detection can be conducted by classifying some specific regions in an image as containing damage or not. For example, Cha et al. [13] used a convolution neural network (CNN) to classify

We thank the National Science Foundation and Amazon Web Services for support from the grant IIS-1741345.

IEEE/ACM ASONAM 2018, August 28-31, 2018, Barcelona, Spain
978-1-5386-6051-5/18/$31.00

small image regions (with 256×256 pixel resolutions) as containing concrete crack damage or not. Maeda et al. [14] used a state-of-the-art object detection approach, called Single Shot MultiBox Detector (SSD) [16], to detect several types of road damage. Image segmentation has been used in [15] to detect building damage based on high resolution aerial images.

Regardless of the method used, object detection or image segmentation, existing approaches for localizing damage first identify objects (i.e., potential damage regions) and subsequently classify the objects as *damage* (sometimes, *severe* or *mild*) or *no damage*. Thus, there is a conceptual mismatch in the way existing approaches are used, given that damage is generally regarded as a high-level concept rather than a well-defined object. By first identifying objects and then assigning discrete hard-labels to them, existing approaches produce a clear-cut boundary for the damaged areas, although a smooth boundary would be more appropriate.

Contributions: Inspired by the technique called Class Activation Mapping (CAM) [17], we propose a novel approach, called Damage Detection Map (DDM), to generate a smooth damage heatmap for an image. Our approach adopts the gradient-weighted CAM [18] technique to localize the area in an image which contributes to the damage class. Based on the damage heatmap, we also propose a new quantitative measure, called Damage Assessment Value (DAV), to quantify the severity of damage on a continuous scale. One advantage of the proposed approach is that it only requires annotators to label images as having damage or no damage, as opposed to requiring annotators to localize the damage. Thus, our approach makes the disaster damage localization possible, and extends the use of social media images in disaster assessment.

The rest of this paper is organized as follows: we describe the proposed approaches for generating DDM heatmaps, and computing DAV scores in Section II. We describe the experimental setup and results in Section III. We discuss related work in Section IV, and conclude the paper in Section V.

II. Proposed Approach

Our approach generates a Damage Detection Map, which visualizes the damage area for a given image, and a score DAV, which quantifies the severity of the damage. The main components of our approach, shown in Fig. 1, are the following: 1) a CNN that classifies images into two classes, *damage* or *no damage*; 2) a class activation mapping, which generates the DDM map by weighting the last convolutional layer of the CNN model; 3) finally, the damage severity score computed by averaging the values in the map. The details for the three components of our approach are provided in what follows.

A. Convolutional Neural Networks

Convolutional neural networks [19] have been used successfully for many image analysis tasks [20], [21]. The ImageNet annual competition (where a dataset with 1.2 million images in 1000 categories is provided to participants) has led to several popular architectures, including AlexNet [22], VGG19 [23], ResNet [24] and Inception [25]. We choose VGG19 as the architecture for our CNN model, as VGG19 has good classification accuracy and it is relatively simpler compared to ResNet and Inception. Furthermore, the pre-trained model are available for VGG19 and it is easy to fine tune them for different classification problems.

VGG19 [23] contains 16 convolutional layers (with 5 pooling layers) and 3 fully connected layers. Each convolution layer is equipped with a non-linear ReLU activation [22]. The convolutional layers can be seen as feature extraction layers, where each successive layer detects predictive image features (i.e., image fragments that correspond to edges, corners, textures, etc.) at a more abstract level than the previous layer.

As can be seen in Fig. 1, the size of the input to the convolutional layers is reduced by a factor of 2 through max-pooling layers, but not all convolution layers are followed by a max-pooling layer. After every max-pooling layer, the width of the convolution layer (i.e., number of filters used) increases by a factor of 2. After the last max-pooling layer, there are two fully connected layers with dimension 4096, and another fully connected layer whose neurons correspond to the categories to be assigned to the input image. The last layer of the standard VGG19 model has dimension 1000 because VGG19 was originally trained on a dataset with 1000 categories. However, as we are interested in using VGG19 to classify images in two categories, *damage* and *no damage*, we change the dimension of the last fully connected layer from 1000 to 2. Overall, the model includes more than 130 million parameters, and it takes a significant amount of time (and a large number of images) to train it accurately [23]. However, the model parameters are highly transferable to other image classification problems [26]. Thus, to avoid the need for a large number of images, we initialize our model with the pre-trained VGG19 model (except for the last fully connected layer), and fine tune it using disaster-related images. More specifically, given a training image x with label y represented as a one-hot vector (e.g., if the label is *damage*, then $y = [1, 0]$, otherwise $y = [0, 1]$), all the parameters θ will be updated by:

$$\theta \longleftarrow \theta - \mu \frac{\partial \mathcal{L}(y, \text{CNN}(x))}{\partial \theta}, \tag{1}$$

where μ is learning rate, $\mathcal{L}$ is the cross-entropy loss, and $\text{CNN}(x)$ is the output of the CNN given input x.

B. Damage Detection Map

Our proposed Damage Detection Map (DDM) is inspired by the Gradient-weighted Class Activation Mapping (GCAM) [18]. In a general classification problem, for an input image and a trained CNN model, GCAM makes use of the gradients of a target category to compute a category-specific weight for each feature map of a convolution layer. The weights are used to aggregate the feature maps of the final convolutional layer, under the assumption that the last level captures the best trade-off between high-level semantic features and spatial information. The resulting maps can be used to identify the discriminative regions for the target category (which explain the CNN model's prediction), and implicitly to localize the

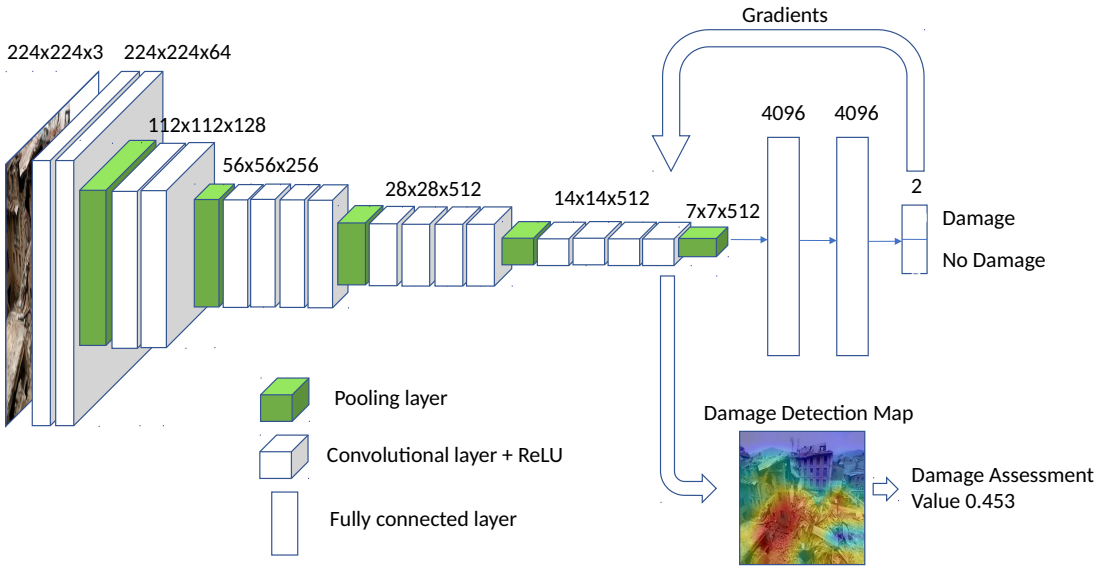


Fig. 1: Overview of the proposed approach.

category in the input image. Thus, GCAM can be seen as a weakly supervised approach, which can localize a category in an image based only on global image labels [18]. Furthermore, the GCAM localized categories or objects (shown using heatmaps) have soft boundaries, and can be used to gain both insight and trust into the model. This makes GCAM particularly attractive for the problem of localizing damage in disaster images, assuming that only coarse labeling of images as *damage* or *not damage* is available for training. The heatmaps showing categories of interest using soft-boundaries are very appropriate for localizing damage, as damage boundaries are inherently soft. Moreover, the heatmaps that explain the model's predictions can be used to gain the trust of disaster management teams, and thus increase the usability of social media images in disaster response and recovery.

Given the above introduction to GCAM and motivation for use in the context of disaster damage localization, we formally describe the GCAM approach [18] in remaining of this subsection. Consider a 14×14 matrix S, such that

$$s_{i,j} = \text{ReLU}(\sum_k w_k f^k_{(i,j)}) \tag{2}$$

where w_k is a gradient-based weight parameter (defined in Equation 3) corresponding to each feature map f^k in the last convolutional layer (of dimension $14 \times 14 \times 512$), and $f^k_{(i,j)}$ represents the value at location (i, j) in the k-th feature map, for $i = 1, \ldots, 14$, $j = 1, \ldots, 14$, $k = 1, \ldots, 512$. The ReLU function is used to cancel the effect of the negative values, while emphasizing the effect of the positive values. Given an input image x and the output of the CNN for the damage class y_D (just before the softmax function is applied), the weights w_k are determined as the sum of the gradients of output y_D with respect to $f^k_{(i,j)}$, for all i, j. Specifically:

$$w_k = \frac{1}{14 \times 14} \sum_{i,j} \frac{\partial y_D}{\partial f^k_{(i,j)}} \tag{3}$$

The feature maps f^k in Equations (3) and (2) are in the last convolutional layer, as this layer generally shows a good trade-off between high-level features and spatial information in the original image. Our final goal is to localize damage in disaster images, in other words to find regions that are discriminative for the damage class. Intuitively, the gradient with respect to a neuron in the final convolutional layer represents the contribution of the neuron to the class label. The procedure described above produces a 14×14 matrix S (and correspondingly a 14×14 image). Using an interpolation technique, we further resize S to S_C to match the original dimensions of the image. The heatmap showing the S_C values is the final Damage Detection Map.

C. Measuring the damage severity

When a disaster occurs, eyewitnesses of the disaster will produce a huge number of images in a short period of time. An important goal of disaster assessment is to extract and concisely summarize the information contained in the images posted by eyewitnesses. To meet this demand, we propose to use a damage assessment value (DAV) derived from the DDM heatmap to represent the damage severity for each image.

The disaster damage map uses numerical values to measure the intensity of each pixel of the image (the higher the intensity, the more severe the damage), and can be represented as a heatmap. We take the average over all the numerical

TABLE I: Social media image dataset consisting of images from four disaster events. Images are labeled as *Severe*, *Mild*, or *None*. The number of images in each class, and the total number of images for each disaster are shown.

Disaster	Severe	Mild	None	Total
Nepal Earthquake	5303	1767	11226	18296
Ecuador Earthquake	785	83	886	1754
Ruby Typhoon	76	325	463	864
Matthew Hurricane	97	89	130	316

values in the heatmap of a given image, and use the resulting value as an overall score for the severity of the damage. Formally, we define the damage assessment value (DAV) as:

$$DAV = \frac{1}{14 \times 14} \sum_{i,j} s_{i,j} \tag{4}$$

where $s_{i,j}$ are the elements of the S matrix defined in Equation (2), and 14×14 is the dimension of the matrix S.

III. Experimental Results

We perform a series of experiments to evaluate the performance of our proposed method. The experiments are designed to answer the following questions: (i) can the Damage Detection Map accurately locate the damage areas, (ii) can the DAV score provide a reliable measure for damage severity.

A. Experimental Setting

1) Datasets: We used two datasets in our experiments. The first dataset is assembled using Google image search engine. Specifically, we performed two searches: first, we used 'nepal', 'building' and 'damage' as keywords, and crawled 308 *damage* images from the result; second, we used 'nepal' and 'city' as keywords, and crawled 311 *no damage* images. The keyword 'nepal' was used in both searches to ensure that the two sets of images have similar scene, while the keywords 'building damage' and 'city' were used to bias the search towards *damage* and *no damage* images, respectively. We show some sample *damage* and *no damage* images from the Google dataset in Fig. 2 in the first and third rows, respectively.

The second dataset used in our evaluation was previously used in [12], and consists of social media images posted during four different disaster events: Nepal Earthquake, Ecuador Earthquake, Ruby Typhoon, and Matthew Hurricane. For each disaster, the dataset contains images in three categories, representing three levels of damage: severe, mild, and none. Table I provides the class distribution for each disaster dataset.

In our experiments, each dataset is randomly split into a training set (80%) and a test set (20%). We will use the proposed approach to learn a CNN that discriminates between *damage* (including *severe* and *mild*) and *no damage* images, while also localizing the damage and identifying image features discriminative for the *damage* class.

2) Hyper-parameters: We used TensorFlow's GradientDescentOptimizer to train the model using mini-batch gradient descent on a GeForce GTX 1070 graphic card. Based on preliminary experimentation with the Ecuador Earthquake and Ruby Typhoon datasets, we chose to use a learning rate of 0.001 and a batch size of 32 images in all our experiments. Furthermore, we used the dropout technique with a rate of 0.5 to prevent overfitting. The code for the VGG19 model was adapted from https://github.com/machrisaa/tensorflow-vgg.

B. Damage Detection Map Evaluation

We first use the Google dataset, where the *damage* images are easier to discriminate from the *no damage* images, to evaluate the Damage Detection Map (DDM) approach (intuitively the better the CNN classifier, the better the DDM map). Specifically, we train a CNN model on the Google dataset as described in Section II-A. The training accuracy of the CNN model is 100%, and the test accuracy is 95.5%. Subsequently, we compute the DDM heatmaps as described in Section II-B. The DDM heatmaps corresponding to the sample *damage* and *no damage* images are shown in Fig. 2, in the second and fourth rows, respectively. As can be seen, the regions with high intensity DDM generally correspond to damage.

To answer our first research question, specifically to understand if the DDM can accurately locate the damage areas, we will evaluate the localization capability of DDM by using the Intersection-Over-Union (IOU) measure, which is frequently used to evaluate object detection techniques [27]. The IOU measure is defined as follows:

$$\text{IOU} = \frac{\text{Area of Overlap}}{\text{Area of Union}} \tag{5}$$

where the 'Area of Overlap' and 'Area of Union' are computed with respect to a ground truth image, where damage is manually marked. IOU takes values in $[0, 1]$. If the IOU value for a pair of GCAM damage marked image and the corresponding ground truth image is large, then the marked damage areas in the two images are similar, and thus the GCAM approach performs well in terms of damage localization.

We randomly select 10 images from the Google test dataset for manual annotation, and use the annotations to evaluate damage localization using IOU scores. The scores are computed by comparing the automatically localized damage with the manually marked damage. As damage is not an object, heatmaps with smooth boundaries are preferable to bounding boxes when localizing damage. However, human annotators cannot provide precise heatmaps, unless they have professional knowledge of disaster damage, in which case their annotation would be very expensive. To reduce the cost, generally human annotators will simply mark the regions of an image that contain damage (resulting in a binary included/not included representation). We used the tool LabelMe (https://github.com/wkentaro/labelme) to mark the damage.

To compare heatmaps with images marked by annotators in terms of IOU values, we transform the heatmaps to a binary representation as follows: we determine the maximum value in S_C and use 20% of the maximum value as a cutoff value for including a region in the disaster damage "object" or not [17]. In other words, only the regions in DDM with values larger than the cutoff value will be part of the localized damage. In the resulting transformed image (as well as in the human

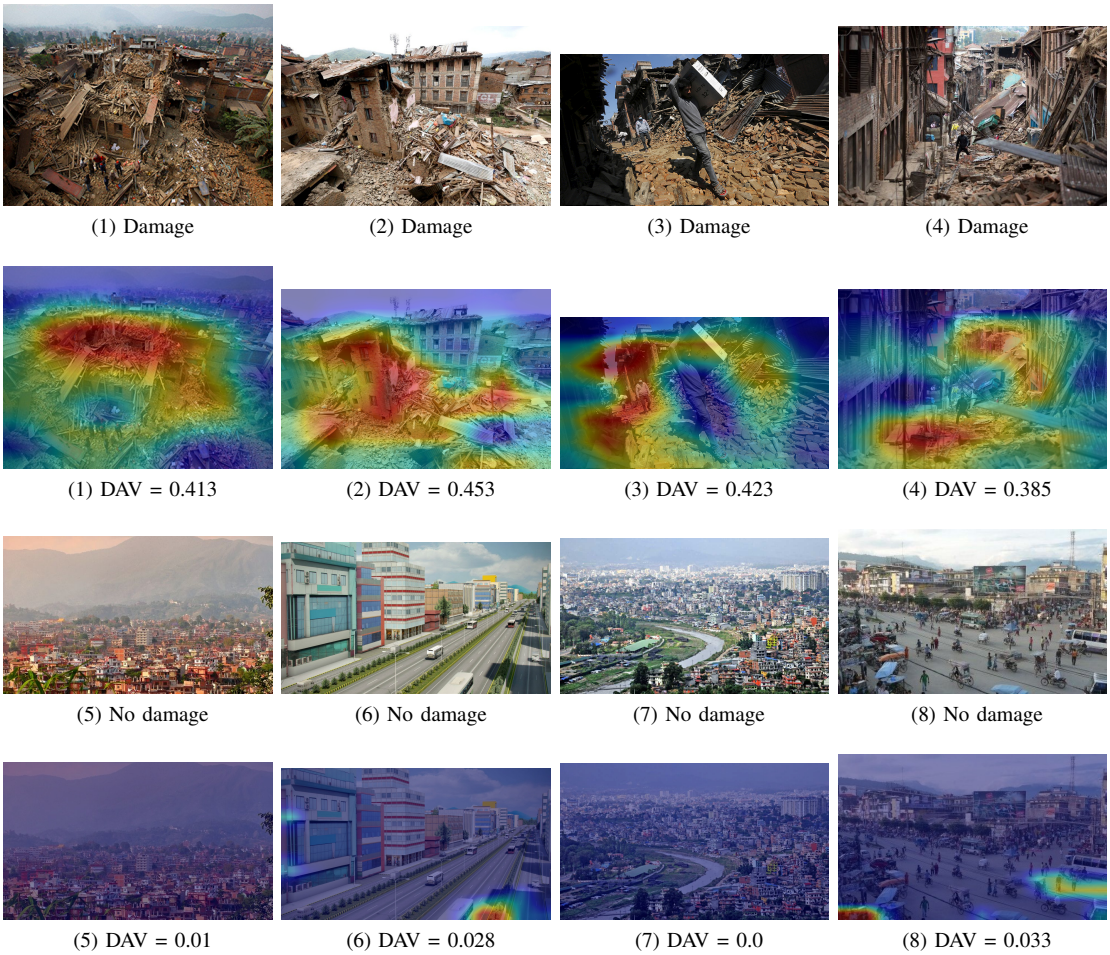


Fig. 2: Examples of images in our dataset, and their corresponding DDM heatmaps and DAV scores. The first row shows examples of images in the *damage* class, followed by their corresponding DDM heatmaps and the DAV scores. Similarly, the third row shows examples of images in the *no damage* class, followed by their corresponding DDM heatmaps and DAV scores.

annotated images), the damage pixels have value 255, while no damage pixels have value 0. We show the result of the IOU evaluation on the sample consisting of 10 Google images, by comparison with annotations from two independent annotators A and B, in Table II. To understand the difficulty of the task of identifying damage in a picture, we also compute IOU values that show the agreement between annotators A and B.

Conventionally, if the IOU value corresponding to a detected object (marked with a bounding box) is larger than 0.5, the detection/localization of that object is considered to be correct [27]. However, given that the damage is not an object but a concept, we find that the 0.5 threshold is too strict in the context of detecting and localizing damage, especially as the average IOU agreement between annotators is 0.610. If we consider an IOU score of 0.4 as detection, then the GCAM-based approach detects 50% of the damage marked by annotator A, and 90% of the damage marked by annotator B. To better understand the results, in Fig. 3, we show the annotations for three sample images, specifically, image 5 for which the IOU agreement with both annotators is above 0.4, image 7 for which the IOU agreement with both annotators is smaller than 0.4, and image 1, for which the agreement with Annotator A is below 0.4 and the agreement with the Annotator B is above 0.4. As can be seen, the damage areas marked based on the DDM heatmaps look accurate overall, and in some cases better than the damage areas marked by the annotators, which confirms that identifying damage is a subjective task, and that DDM can be useful in localizing damage and providing visualizations that can help responders gain trust in the annotations produced by deep learning approaches.

C. Damage Assessment Value Evaluation

We calculate the Damage Assessment Value (DAV) as described in Section II-C. The DAV values for the sample Google images in Fig. 2 are shown below the corresponding heatmap images. As can be seen, images that present a more severe damage scene have higher DAV values, while images with no damage have DAV values very close to zero.

We also use the disaster datasets shown in Table I to further evaluate the DAV values. For each disaster, we use the corresponding training set to fine-tune the CNN model to that

TABLE II: IOU for 10 Google images. Annotation A and Annotation B are provided independently by two annotators. The image heatmaps were transformed to a binary representation by using a threshold equal to 20% of the max value in the DDM

	Image	1	2	3	4	5	6	7	8	9	10	Average
IOU	Annotation A versus DDM	0.334	0.401	0.568	0.360	0.474	0.270	0.349	0.156	0.418	0.477	0.380 $\pm$ 0.116
	Annotation B versus DDM	0.444	0.623	0.633	0.473	0.583	0.445	0.258	0.440	0.616	0.658	0.517 $\pm$ 0.126
	Annotation A versus B	0.677	0.541	0.744	0.681	0.695	0.546	0.673	0.237	0.621	0.689	0.610 $\pm$ 0.146

(a) Original 1 (b) DDM (c) DDM+Threshold (d) Annotator A (e) Annotator B

(a) Original 5 (b) DDM (c) DDM+Threshold (d) Annotator A (e) Annotator B

(a) Original 7 (b) DDM (c) DDM+Threshold (d) Annotator A (e) Annotator B

Fig. 3: Damage Detection Map (DDM) versus human annotations for images 1, 5 and 7 in Table II. (a) Original Google image; (b) DDM damage heatmap; (c) Binary representation of the DDM computed with a threshold whose value is 20% of the max value in the DDM heatmap; (d) Damage marked by Annotator A; (e) Damage marked by Annotator B.

specific disaster. For this purpose, we combined the original *severe* and *mild* damage classes into a single *damage* class, while the original *none* class label is used as *no damage* class. We used the CNN model of each disaster to produce DDM heatmaps for all test images from that disaster. Subsequently, we used the heatmaps to generate DAV values for each test image, and generated DAV density plots for the *damage* versus *no damage* classes, shown in Fig. 5. As can be seen in the figure, the density plots show clearly differentiable patterns between the two classes, with the *damage* class exhibiting higher values overall, as expected. However, there is also overlap in terms of DAV values for *damage* and *no damage* images. This can be partly explained by the noisiness of the dataset, and the difficulty of the task of separating images in damage severity classes, as also noted in [12]. To illustrate this claim, in Fig. 4, we show several images that seem to be mislabeled in the original dataset.

D. Classification using DAV values

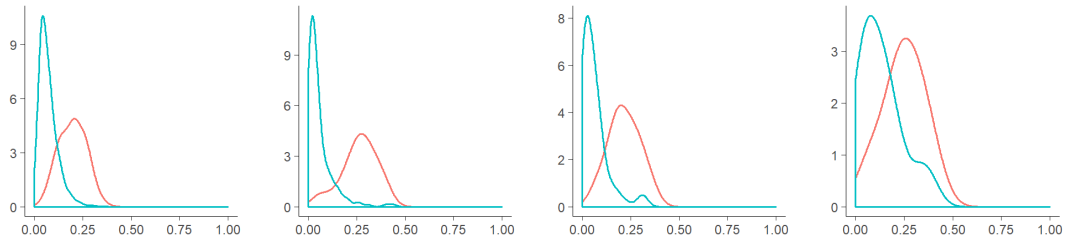
In this section, we study the usefulness of the DAV values in classifying images into several damage categories. Specifically, we consider the *severe*, *mild*, and *none* categories, as the images in the disasters used in our study are already labeled with these categories. Using the training data, we perform a grid-search to find two threshold values for DAV, denoted by c_1 and c_2, which minimize the classification error on the training data. Using these thresholds, we design a simple classifier as follows: test images with DAV values smaller than c_1 will be classified as *none*, those with DAV values in between c_1 and c_2 will be classified as *mild*, and finally those with DAV values greater than c_2 will be classified as *severe*. The classification results of our simple classifier (averaged over five independent runs), together with the average results of a three-class CNN model (trained on the corresponding training set of each disaster in each run), are reported in Table III.The classification results of the simple classifier based on DAV are similar to those of the CNN model for Ecuador Earthquake and Matthew Hurricane (i.e., the results are not statistically different based on a t-test with $p \leq 0.05$). While the CNN accuracy looks better overall, the results indicate that DAV can capture severity damage (on a continuous scale), and it can help produce simple and interpretable classifiers.

IV. Related Work

Social media data has been shown to have significant value in disaster response [29]–[31]. Many machine learning approaches [6], [8], [32], including deep learning approaches [28], [33], have been proposed and used to help identify and

Mild, DAV=0.328 None, DAV=0.307 None, DAV=0.435 Mild, DAV=0.375 Severe, DAV=0.376

Fig. 4: Examples of images mislabeled in the disaster dataset. The original image label and the DAV score are shown.



DAV (Nepal Earthquake) DAV (Ecuador Earthquake) DAV (Ruby Typhoon) DAV (Matthew Hurricane)

Fig. 5: Smoothed density curves for DAV values. The *no damage* class is shown in blue and the *damage* class is shown in red. The graphs are based on the test set of each disaster dataset.

TABLE III: Classification accuracy for each disaster dataset. The first row shows the average accuracy of our simple DAV-based classifiers (over 5 independent runs), together with standard deviation. The second row shows the average accuracy of the three-class CNN models. The third row shows the results published in [28], which used the same datasets.

Method	Nepal Earthquake	Ecuador Earthquake	Ruby Typhoon	Matthew Hurricane
DAV	0.801±0.013	0.886±0.018	0.793±0.017	0.533±0.043
VGG19	0.851±0.003	0.901±0.013	0.852±0.018	0.603±0.151
VGG16 [28]	0.840	0.870	0.810	0.740

prioritize useful *textual* information (e.g., tweets) in social media. Some works have focused specifically on identifying situational awareness information [34], [35], including information related to damage assessment [7], [9], [36], [37].

Despite the extensive use of machine learning tools for analyzing social media text data posted during disaster events, there is not much work on analyzing social media images posted by eyewitnesses of a disaster. One pioneering work in this area [12], used trained CNN models, specifically, VGG16 networks fine-tuned on the disaster image datasets that are also used in our work (see Table I), and showed that the CNN models perform better than standard techniques based on bags-of-visual-words. Our CNN results, using VGG19 fine-tuned on the same disaster image datasets, are similar to those reported in [12], except for Matthew Hurricane, for which the dataset is relatively small and the model can't be trained well. However, as opposed to [12], where the focus is on classifying images into three damage categories, we go one step further and use the Grad-CAM approach [18] to localize damage, with the goal of improving the trust of the response teams in the predictions of the model. Furthermore, we produce a continuous severity score, as a way to quantify damage.

Other prior works focused on image-based disaster damage assessment use aerial or satellite images, e.g. [1], [2]. Compared to such works, which use more expensive imagery, we focus on the use of social media images, which are readily available during disasters, together with interpretability approaches, i.e., Grad-CAM, to produce a damage map and a damage severity score for each image.

Similar to us, Nia and Mori [38] use ground-level images collected using Google to assess building damage. Their model consists of three different CNN networks (fine-tuned with raw images, color-masked or binary-masked images, respectively) to extract features predictive of damage. Subsequently, a regression model is used with the extracted features to predict the severity of the damage on a continuous scale. Compared to [38], we use the features identified at the last convolutional layer of the CNN network to build a detection map, and use the map to produce a numeric damage severity score. While CAM-type approaches have been used to explain model predictions in many other application domains, to the best of our knowledge, they have not been used to locate damage and assess damage severity in prior work.

V. Conclusion

Given the large number of social media images posted by eyewitnesses of disasters, we proposed an approach for detecting and localizing disaster damage at low cost. Our approach is built on top of a fine-tuned VGG19 model, and utilizes the Grad-CAM approach to produce a DDM heatmap. Furthermore, the DDM is used to calculate a DAV score for each image. This scoring is performed on a continuous scale

and can be used to assess the severity of the damage. The DAV score, together with the DDM heatmap, can be used to identify and prioritize useful information for disaster response, while providing visual explanations for the suggestions made to increase the trust in the computational models. Quantitative and qualitative evaluations of DDM and DAV components show the feasibility of our proposed approach.

As part of future work, we will compare the Grad-CAM approach used to produce the damage heatmaps with its newer variant Grad-CAM++ [39], whose authors claim to produce heatmaps which better cover the objects of interest as opposed to just identifying the most prominent object features. To understand the robustness of the approach, we will test it on datasets with different types of damage [40], and different resolutions. Also of interest is the applicability of the approach to estimate the global damage produced by a disaster, based on aggregating the DAV values from individual images. Finally, geo-tagging images would enable disaster response teams not only to identify damage, but also to find its physical location.

REFERENCES

[1] S. Xie, J. Duan, S. Liu, Q. Dai, W. Liu, Y. Ma, R. Guo, and C. Ma, "Crowdsourcing rapid assessment of collapsed buildings early after the earthquake based on aerial remote sensing image: A case study of yushu earthquake," *Remote Sensing*, vol. 8, no. 9, p. 759, 2016.

[2] L. Gueguen and R. Hamid, "Large-scale damage detection using satellite imagery," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1321–1328.

[3] Y. Fan, Q. Wen, W. Wang, P. Wang, L. Li, and P. Zhang, "Quantifying disaster physical damage using remote sensing dataa technical work flow and case study of the 2014 ludian earthquake in china," *International Journal of Disaster Risk Science*, vol. 8, no. 4, pp. 471–488, 2017.

[4] N. Attari, F. Ofli, M. Awad, J. Lucas, and S. Chawla, "Nazr-cnn: Fine-grained classification of uav imagery for damage assessment," in *Data Science and Advanced Analytics (DSAA), 2017 IEEE International Conference on*. IEEE, 2017, pp. 50–59.

[5] X. Guan and C. Chen, "Using social media data to understand and assess disasters," *Natural hazards*, vol. 74, no. 2, pp. 837–850, 2014.

[6] M. Imran, C. Castillo, F. Diaz, and S. Vieweg, "Processing social media messages in mass emergency: A survey," *ACM Computing Surveys (CSUR)*, vol. 47, no. 4, p. 67, 2015.

[7] Y. Kryvasheyeu, H. Chen, N. Obradovich, E. Moro, P. Van Hentenryck, J. Fowler, and M. Cebrian, "Rapid assessment of disaster damage using social media activity," *Science advances*, vol. 2, no. 3, 2016.

[8] H. Li, D. Caragea, C. Caragea, and N. Herndon, "Disaster response aided by tweet classification with a domain adaptation approach," *Journal of Contingencies and Crisis Management (JCCM), Special Issue on HCI in Critical Systems.*, vol. 26, no. 1, pp. 16–27, 2017.

[9] F. Yuan and R. Liu, "Feasibility study of using crowdsourcing to identify critical affected areas for rapid damage assessment: Hurricane matthew case study," *International Journal of Disaster Risk Reduction*, 2018.

[10] M. Bica, L. Palen, and C. Bopp, "Visual representations of disaster," in *Proc. of the 2017 ACM CSCW*. NY, USA: ACM, 2017, pp. 1262–1276.

[11] F. Alam, M. Imran, and F. Ofli, "Image4act: Online social media image processing for disaster response," in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*. ACM, 2017, pp. 601–604.

[12] D. T. Nguyen, F. Ofli, M. Imran, and P. Mitra, "Damage assessment from social media imagery data during disasters," in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*. ACM, 2017, pp. 569–576.

[13] Y.-J. Cha, W. Choi, and O. Büyüköztürk, "Deep learning-based crack damage detection using convolutional neural networks," *Computer-Aided Civil and Infrastructure Engineering*, vol. 32, no. 5, pp. 361–378, 2017.

[14] H. Maeda, Y. Sekimoto, T. Seto, T. Kashiyama, and H. Omata, "Road damage detection using deep neural networks with images captured through a smartphone," *arXiv preprint arXiv:1801.09454*, 2018.

[15] A. Vetrivel, M. Gerke, N. Kerle, F. Nex, and G. Vosselman, "Disaster damage detection through synergistic use of deep learning and 3d point cloud features derived from very high resolution oblique aerial images, and multiple-kernel-learning," *ISPRS journal of photogrammetry and remote sensing*, vol. 140, pp. 45–59, 2018.

[16] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg, "SSD: single shot multibox detector," in *In European conference on computer vision*. Springer, 2016, p. 2137.

[17] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2921–2929.

[18] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[19] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, 1989.

[20] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[21] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[25] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[26] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Advances in neural information processing systems*, 2014, pp. 3320–3328.

[27] M. Everingham, L. Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (voc) challenge," *Int. J. Comput. Vision*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[28] D. T. Nguyen, K. Al-Mannai, S. R. Joty, H. Sajjad, M. Imran, and P. Mitra, "Rapid classification of crisis-related data on social networks using convolutional neural networks," *CoRR*, vol. abs/1608.03902, 2016.

[29] C. Castillo, *Big Crisis Data: Social Media in Disasters and Time-Critical Situations*. Cambridge University Press, 2016.

[30] P. Meier, *Digital Humanitarians: How Big Data Is Changing the Face of Humanitarian Response*. FL, USA: CRC Press, Inc., 2015.

[31] L. Palen and K. M. Anderson, "Crisis informatics-new data for extraordinary times," *Science*, vol. 353, no. 6296, pp. 224–225, 2016.

[32] Z. Ashktorab, C. Brown, M. Nandi, and A. Culotta, "Tweedr: Mining twitter to inform disaster response," *Proc. of ISCRAM*, 2014.

[33] C. Caragea, A. Silvescu, and A. H. Tapia, "Identifying informative messages in disasters using convolutional neural networks," in *Proc. of the ISCRAM, Brazil*, 2016.

[34] A. Sen, K. Rudra, and S. Ghosh, "Extracting situational awareness from microblogs during disaster events," in *7th Int. Conf. on Communication Systems and Networks (COMSNETS)*. IEEE, 2015, pp. 1–6.

[35] Q. Huang and Y. Xiao, "Geographic situational awareness: mining tweets for disaster preparedness, emergency response, impact, and recovery," *ISPRS Int. Journal of Geo-Information*, vol. 4, no. 3, 2015.

[36] B. Resch, F. Usländer, and C. Havas, "Combining machine-learning topic models and spatiotemporal analysis of social media data for disaster footprint and damage assessment," *Cartography and Geographic Information Science*, pp. 1–15, 2017.

[37] M. Enenkel, S. M. Saenz, D. S. Dookie, L. Braman, N. Obradovich, and Y. Kryvasheyeu, "Social media data analysis and feedback for advanced disaster risk management," *CoRR*, vol. abs/1802.02631, 2018.

[38] K. R. Nia and G. Mori, "Building damage assessment using deep learning and ground-level image data," in *14th Conference on Computer and Robot Vision (CRV)*. IEEE, 2017, pp. 95–102.

[39] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," *CoRR*, vol. abs/1710.11063, 2017.

[40] F. Alam, F. Ofli, and M. Imran, "Crisismmd: Multimodal twitter datasets from natural disasters," *arXiv preprint arXiv:1805.00713*, 2018.